Randomized chi-square goodness of fit test and maximum test (Statistical Inference of Records and Related Statistics)

田中, 俊彦; 赤平, 昌文

数理解析研究所講究録 1439 巻 2005 年 128-150

Randomized chi-square goodness of fit test and maximum test

早大・理工　田中 俊彦　(Toshihiko Tanaka)
筑波大・数理　赤平 昌文　(Masafumi Akahira)

1 はじめに

一般に, 想定された分布型がデータに当てはまっているかどうかを検定することを適合度検定 (goodness of fit test) という. この検定の特定の分布型の特徴を用いない場合に Pearson のカイ 2 乗適合度検定, 用いる場合に Kolmogorov-Smirnov 検定がよく知られている. 本論の第 2 節では, Greenwood and Nikulin[GN96], Lehmann[Leh99] に従って, カイ 2 乗 (χ^2) 適合度検定において χ^2 統計量の漸近的な性質について論じる. 次に, 第 3 節では, χ^2 統計量の代わりに, より簡便な統計量として最大値統計量によるマックス (max) 検定を考える. そして, マックス検定と χ^2 適合度検定の検出力の比較を行う. さらに第 4 節では, χ^2 統計量は離散値をとるため与えられた水準をもつ χ^2 適合度検定を行うことは一般にはできないので, 離散的な値をとる χ^2 統計量に連続な値をとる確率変数を加えることによって, 連続な値をとることができるようにしたランダム χ^2 統計量による検定を考える. そして, 従来の χ^2 適合度検定とランダム χ^2 適合度検定の比較を行う.

2 カイ 2 乗適合度検定とマックス検定

まず, 確率ベクトル $(X_1, \cdots, X_k)$ が確率関数

$$\begin{aligned} P(\boldsymbol{X}=\boldsymbol{x}) &= P(X_1=x_1, \cdots, X_k=x_k) \\ &= \frac{n!}{x_1!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \end{aligned}$$

を持つ多項分布 $M_k(n; p_1, \cdots, p_k)$ に従っているとする. ただし, $\boldsymbol{x} := (x_1, \cdots, x_k)$ とし, $x_1, \cdots, x_k$ は非負の整数で, $x_1 + \cdots + x_k = n$ とし, また $0 < p_i < 1$ $(i = 1, \cdots, k)$, $p_1 + \cdots + p_k = 1$ とする. このとき

$$E\left(\frac{X_i}{n}\right) = p_i \quad (i = 1, \cdots, k)$$

となり, 多変量の場合の中心極限定理から

$$\left(\sqrt{n}\left(\frac{X_1}{n} - p_1\right), \cdots, \sqrt{n}\left(\frac{X_k}{n} - p_k\right)\right) \xrightarrow{L} N(\mathbf{0}, \Sigma) \tag{1}$$

になる. ここで, 一般に確率ベクトル $\boldsymbol{Z}_n$, 確率分布 F について, "$\boldsymbol{Z}_n \xrightarrow{L} F$" は $\boldsymbol{Z}_n$ が F に従う確率ベクトル $\boldsymbol{Z}$ に法則収束することを意味し, また記号で $\boldsymbol{Z}_n \xrightarrow{L} \boldsymbol{Z}$ とも表す. そして

$\Sigma = (\sigma_{ij})$ は $(\sqrt{n}\,((X_1/n) - p_1), \cdots, \sqrt{n}\,((X_k/n) - p_k))$ の分散共分散行列で,

$$\sigma_{ij} = \begin{cases} p_i(1-p_i) & (i=j), \\ -p_ip_j & (i \neq j) \end{cases}$$

となる. また, $\Sigma^{-1} =: A = (a_{ij})$ とすると

$$a_{ij} = \begin{cases} \dfrac{1}{p_i} + \dfrac{1}{p_k} & (i=j), \\ \dfrac{1}{p_k} & (i \neq j) \end{cases} \tag{2}$$

となる. そして, $0 < p_i^{(0)} < 1\ (i = 1, \cdots k)$, $\sum_{i=1}^k p_i^{(0)} = 1$ として, 仮説

$$H : (p_1, \cdots, p_k) = (p_1^{(0)}, \cdots, p_k^{(0)}) \tag{3}$$

の検定問題を考え, その検定統計量として Pearson のカイ 2 乗 (χ^2) 統計量

$$\chi^2 := \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \tag{4}$$

を用いる. この χ^2 統計量が離散の値をとることに注意. 特に

$$p_1^{(0)} = \cdots = p_k^{(0)} = \frac{1}{k}$$

とすれば, (4) は

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} = (k-1)n - \frac{2k}{n}\sum_{i>j} X_iX_j \tag{5}$$

となるので, (5) より χ^2 の値は区間 $[0, (k-1)n]$ にあり, $2k/n$ の倍数だけ値が離れていることがわかる.

そこで, 仮説 H に対する χ^2 適合度検定は, 有意水準 $\alpha(0 < \alpha < 1)$ について棄却限界値 c_α を定めて, c_α と χ^2 統計量の実現値を比較すればよい, すなわち

$$\chi^2 \geq c_\alpha \text{ならば仮説 } H \text{ を棄却},$$
$$\chi^2 < c_\alpha \text{ならば仮説 } H \text{ を受容}$$

とすればよい. しかし, χ^2 統計量は離散的な値をとるので, 与えられた水準 α に対し, 等式

$$P_H(\chi^2 \geq c_\alpha) = \alpha \tag{6}$$

が成立するような棄却限界値 c_α は必ずしも存在しないことに注意する必要がある. そのため, (6) が成立するような c_α がないときには, 条件

$$l_\alpha < u_\alpha,\quad P_H(\chi^2 = l_\alpha) > 0,\quad P_H(\chi^2 = u_\alpha) > 0,$$
$$P_H(l_\alpha < \chi^2 < u_\alpha) = 0,\quad P_H(\chi^2 \geq l_\alpha) > \alpha > P_H(\chi^2 \geq u_\alpha) > 0$$

をみたす上側棄却限界値 u_α と下側棄却限界値 l_α を一意的に求めることができる. そして, もし,

$$P_H(\chi^2 \geq c_\alpha) \leq \alpha$$

をみたす c_α の最小値とする, すなわち $c_\alpha = u_\alpha$ とすると, (6) は $c_\alpha = u_\alpha = l_\alpha$ の場合に対応する. しかし, χ^2 統計量の正確な確率分布は n が増えると複雑になるので, 確率 $P_H(\chi^2 \geq c_\alpha)$ の近似を考える. 実際, χ^2_{k-1} を自由度 $k-1$ の χ^2 分布 (χ^2_{k-1} 分布) に従う確率変数とすれば, $n \to \infty$ のとき χ^2 は χ^2_{k-1} に法則収束するので, 十分大きい n について

$$P_H(\chi^2 \geq x) \approx P(\chi^2_{k-1} \geq x)$$

となり, 棄却限界値 c_α は漸近的に χ^2_{k-1} 分布の上側 100α%点 $\chi^2_\alpha(k-1)$ となる ([Leh99]).
さて, χ^2 適合度検定の検出力は, 漸近的に

$$P\left(\chi^2_{k-1}(\lambda) > c_\alpha\right)$$

となるが, この確率を直接求めるのは困難なので, k が大きいときに Edgeworth 展開による近似を用いる. 実際, まず

$$U := \frac{\chi^2_{k-1}(\lambda) - (k-1+\lambda)}{\sqrt{2(k-1+2\lambda)}}$$

とおくと, $E(U) = 0,\ V(U) = 1$ で, U の 3 次, 4 次キュムラントはそれぞれ

$$\kappa_{3U} = \frac{2\sqrt{2}(k-1+3\lambda)}{(k-1+2\lambda)^{3/2}},\ \kappa_{4U} = \frac{12(k-1+4\lambda)}{(k-1+2\lambda)^2}$$

となる (Torigoe[To96]). このとき, Edgeworth 展開より

$$\begin{aligned} P\left(\chi^2_{k-1}(\lambda) > c_\alpha\right) &= 1 - P\left(\chi^2_{k-1}(\lambda) \leq c_\alpha\right) \\ &= 1 - P\left(U \leq \frac{c_\alpha - (k-1+\lambda)}{\sqrt{2(k-1+2\lambda)}}\right) \\ &\approx 1 - \left[\Phi(u) - \phi(u)\left\{\frac{1}{6}\kappa_{3U}H_2(u) + \frac{1}{24}\kappa_{4U}H_3(u) + \frac{1}{72}\kappa_{3U}^2 H_5(u)\right\}\right] \end{aligned}$$

となる. ここで,

$$u := \frac{c_\alpha - (k-1+\lambda)}{\sqrt{2(k-1+2\lambda)}}$$

とし, $\Phi(u)$ を $N(0,1)$ の分布関数, $\phi(u)$ を $N(0,1)$ の密度関数, $H_j(u)$ を j 次のエルミート多項式 $\{(-d/du)^j\phi(u)\}/\phi(u)\ (j = 1, 2, \cdots)$ とする. 特に

$$H_2(u) = u^2 - 1,\ H_3(u) = u^3 - 3u,\ H_5(u) = u^5 - 10u^3 + 15u$$

になる.

次に, (3) の仮説 $H:(p_1,\cdots,p_k)=(p_1^{(0)},\cdots,p_k^{(0)})$ の検定において, χ^2 統計量の代わりに統計量として

$$T:=\max_{1\le i\le k} X_i$$

という最大値統計量を考える. この統計量による検定をマックス (max) 検定と呼ぶことにし, まず, この検定の棄却限界値 c を求めよう. いま,

$$P(T>c)=P\left(\max_{1\le i\le k} Xi>c\right)=1-P(X_1\le c,X_2\le c,\cdots,X_k\le c) \tag{7}$$

となり, 最後の式の確率の値を次の定理を用いて求める.

定理 1 (Levin[Lev81]) 確率ベクトル $(X_1,\cdots,X_k)$ を多項分布 $M_k(n;p_1,\cdots,p_k)$ に従うとし, $a_1,\cdots,a_k$ を非負の整数とする. このとき, 任意の正数 $s>0$ について

$$P(X_1\le a_1,\cdots,X_k\le a_k)=\frac{n!}{s^n e^{-s}}\left\{\prod_{i=1}^{k}P(U_i\le x_i)\right\}P(W=n)$$

が成り立つ. ただし, $U_1,\cdots,U_k$ は独立で, 各 U_i はポアソン分布 $Po(sp_i)$ に従う確率変数とし, また, $Y_1,\cdots,Y_k$ が独立で, 各 Y_i が値域 $\{0,1,\cdots,a_i\}$ をもつ切断ポアソン分布 $TPo(sp_i;a_i)$ に従う確率変数とし, $W:=\sum_{i=1}^{k}Y_i$ とする.

証明 まず, 確率変数 $U_1,\cdots,U_k$ を独立で, 各 U_i がポアソン分布 $Po(sp_i)$ に従うとすると, ベイズの定理より

$$\begin{aligned}&P\left(U_1\le a_1,\cdots,U_k\le a_k \;\middle|\; \sum_{i=1}^{k}U_i=n\right)\\ &=\frac{P(U_1\le a_1,\cdots,U_k\le a_k)}{P(\sum_{i=1}^{k}U_i=n)}\,P\left(\sum_{i=1}^{k}U_i=n \;\middle|\; U_1\le a_1,\cdots,U_k\le a_k\right)\end{aligned} \tag{8}$$

になる. ここで, $U_1,\cdots,U_k$ が独立であることと $\sum_{i=1}^{k}U_i\sim Po(s)$ となることから, (8) の右辺は

$$\frac{n!}{e^{-s}s^n}\prod_{i=1}^{k}P(U_i\le a_i)\,P\left(\sum_{i=1}^{k}U_i=n \;\middle|\; U_1\le a_1,\cdots,U_k\le a_k\right)$$

となる. このとき, $P\left(U_1\le a_1,\cdots,U_k\le a_k \;\middle|\; \sum_{i=1}^{k}U_i=n\right)$ が多項分布の累積分布関数 (c.d.f.) であることと, 各 $i=1,\cdots,k$ について, $U_i\le a_i$ を与えた U_i の条件付き分布が $TPo(sp_i;a_i)$ になることから定理は証明される. □

また, 後の計算のために $TPo(sp_i;a_i)$ に従う確率変数のモーメントに関する次の定理を挙げる ([Lev81], [JK92]).

定理 2 確率変数 Y を切断ポアソン分布 $TPo(\lambda; m)$ に従うとすると

$$\mu := E(Y) = \lambda\left(1 - \frac{\lambda^m/m!}{\sum_{i=0}^{m}\lambda^i/i!}\right) = \lambda\left(1 - \frac{P(X=m)}{P(X \le m)}\right)$$

である. ただし, 確率変数 $X \sim Po(\lambda)$ とする. また

$$\mu_{(r)} := E[Y(Y-1)\cdots(Y-r+1)],$$
$$m^{(r)} := m(m-1)\cdots(m-r+1)$$

とすると

$$\mu_{(1)} = \mu,$$
$$\mu_{(r+1)} = \lambda\mu_{(r)} - m^{(r)}(\lambda - \mu) \quad r = 1, 2, \cdots$$

である. 特に

$$\mu_2 := V(Y) = \mu_{(2)} + (\mu - \mu^2) = \mu - (m-\mu)(\lambda-\mu),$$
$$\mu_3 := E[(Y-\mu)^3] = \mu_{(3)} + \mu_{(2)}(3-3\mu) + (\mu - 3\mu^2 + 2\mu^3),$$
$$\mu_4 := E[(Y-\mu)^4] = \mu_{(4)} + \mu_{(3)}(6-4\mu) + \mu_{(2)}(7 - 12\mu + 6\mu^2) + (\mu - 4\mu^2 + 6\mu^3 - 3\mu^4)$$

である.

証明 確率変数 Y が $TPo(\lambda; m)$ に従うので, その確率関数は

$$P(Y=y) = \frac{\lambda^y}{y!}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} \quad (y = 0, 1, \cdots, m)$$

であるから

$$\begin{aligned}
\mu = E(Y) &= \sum_{y=0}^{m} y\frac{\lambda^y}{y!}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} \\
&= \left(-\frac{\lambda^{m+1}}{m!} + \lambda\sum_{y=0}^{m}\frac{\lambda^y}{y!}\right)\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} \\
&= -\frac{\lambda^{m+1}}{m!}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} + \lambda \\
&= \lambda\left(1 - \frac{\lambda^m/m!}{\sum_{i=0}^{m}\lambda^i/i!}\right)
\end{aligned}$$

となる. 次に

$$\begin{aligned}
\mu_{(r+1)} &= E[Y(Y-1)\cdots(Y-r)] \\
&= \left\{-\frac{\lambda^{m+r+1}}{m!} - \cdots - \frac{\lambda^{m+1}}{(m-r)!} + \lambda^{r+1}\sum_{y=0}^{m}\frac{\lambda^y}{y!}\right\}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} \\
&= \left\{-\frac{\lambda^{m+r+1}}{m!} - \cdots - \frac{\lambda^{m+1}}{(m-r)!}\right\}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} + \lambda^{r+1}, \\
\mu_{(r)} &= E[Y(Y-1)\cdots(Y-r+1)] \\
&= \left\{-\frac{\lambda^{m+r}}{m!} - \cdots - \frac{\lambda^{m+1}}{(m-r+1)!}\right\}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} + \lambda^{r}
\end{aligned}$$

より

$$\begin{aligned}
&\lambda\mu_{(r)} - m^{(r)}(\lambda-\mu) \\
=&\lambda\left[\left\{-\frac{\lambda^{m+r}}{m!} - \cdots - \frac{\lambda^{m+1}}{(m-r+1)!}\right\}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} + \lambda^{r}\right] \\
&\quad - m(m-1)\cdots(m-r+1)\left[\lambda - \left\{-\frac{\lambda^{m+1}}{m!}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} + \lambda\right\}\right] \\
=&\left\{-\frac{\lambda^{m+r+1}}{m!} - \cdots - \frac{\lambda^{m+2}}{(m-r+1)!}\right\}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} + \lambda^{r+1} - \frac{\lambda^{m+1}}{(m-r)!}\left(\sum_{i=0}^{m}\frac{\lambda^i}{i!}\right)^{-1} \\
=&\mu_{(r+1)}
\end{aligned}$$

となる. □

定理 1 より (7) は,

$$P(T>c) = 1 - \frac{n!}{s^n e^{-s}}\left\{\prod_{i=1}^{k} P(U_i \le c)\right\} P(W=n) \tag{9}$$

となるが, $s=n$ として n が大きいとき Stirling の公式から

$$\frac{n!}{n^n e^{-n}} \approx \sqrt{2\pi n}$$

となる. また, 定理 2 より各 $i = 1,\cdots,k$ について $\mu_i := E(Y_i)$, $\sigma_i^2 := V(Y_i)$, $\mu_{3,i} := E[(Y_i-\mu_i)^3]$, $\mu_{4,i} := E[(Y_i-\mu_i)^4]$ として, k が大きいとき, Edgeworth 展開より

$$P(W=n) \approx f\left(\frac{n-\sum_{i=1}^{k}\mu_i}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}}\right)\frac{1}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}}$$

と近似される. ここで,

$$\gamma_1 := \frac{1}{\sqrt{k}} \frac{\frac{1}{k}\sum_{i=1}^{k}\mu_{3,i}}{(\frac{1}{k}\sum_{i=1}^{k}\sigma_i^2)^{3/2}}, \quad \gamma_2 := \frac{1}{k} \frac{\frac{1}{k}(\sum_{i=1}^{k}\mu_{4,i} - 3\sigma_i^4)}{(\frac{1}{k}\sum_{i=1}^{k}\sigma_i^2)^2}$$

として,

$$\begin{aligned} f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Big\{ 1 &+ \frac{\gamma_1}{6}(x^3 - 3x) + \frac{\gamma_2}{24}(x^4 - 6x^2 + 3) \\ &+ \frac{\gamma_2}{72}(x^6 - 15x^4 + 45x^2 - 15) \Big\} \end{aligned}$$

とする. 従って, (9) は, k が大きいとき

$$\begin{aligned} P(T > c) &\approx 1 - \sqrt{2\pi n} \left\{ \prod_{i=1}^{k} P(U_i \le c) \right\} f\left(\frac{n - \sum_{i=1}^{k}\mu_i}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}} \right) \frac{1}{\sqrt{\sum_{i=1}^{k}\sigma_i^2}} \\ &=: 1 - \widetilde{F}_T(c) \end{aligned} \tag{10}$$

となるから, $0 < \alpha < 1$ について $\widetilde{F}_T(c) = 1 - \alpha$ となるように c を定めれば, c は T の分布の漸近的な上側 $100\alpha\%$ 点になる.

例 1 いま, $n = 1000,\ k = 12,\ \alpha = 0.05$ とし,

$$H : (p_1, \cdots, p_{12}) = (1/12, \cdots, 1/12)$$

とする. このとき, (10) より

$$P_H\left(\max_{1\le i\le 12} X_i > 106 \right) = 0.0611,$$
$$P_H\left(\max_{1\le i\le 12} X_i > 107 \right) = 0.0450$$

となるからここでは棄却限界値 $c = 107$ と定めるのが妥当であるといえる.

次に, マックス検定と χ^2 検定の検出力を比較する. 比較に際して問題となるのが, マックス検定の棄却限界値 c が整数値を取るため, 先の例 1 から分かるように必ずしも有意水準を達成する棄却限界値が得られないことである. そこで, 次のようなランダム化を考える.

まず, 非ランダム検定の観点から, c を非負の整数としてマックス検定の棄却域は $\left\{ \max_{1\le i\le k} X_i \ge c \right\}$ となる. そこで, 水準 α のランダム検定関数

$$\phi(\boldsymbol{X}) = \begin{cases} 1 & (\max_{1\le i\le k} X_i > c + 1), \\ \gamma & (\max_{1\le i\le k} X_i = c + 1), \\ 0 & (\text{その他}) \end{cases}$$

を考える. ただし, $\boldsymbol{X} = (X_1, \cdots, X_k)$ とし, 整数 c, $0 \le \gamma \le 1$ は $E_H(\phi) = \alpha$ となるように定める. しかし, このランダム検定は必ずしも実用的ではないので, 実際には次のように行う. まず, 確率変数 U を一様分布 $U(0,1)$ に従い, $X_1, \cdots, X_k$ とは独立とする. そして

$$S := \max_{1 \le i \le k} X_i + U$$

とおくと, 棄却域を $\{S \ge c+1-\gamma\}$ ともつ非ランダム検定はランダム検定 ϕ と同等になる.

このように最大値統計量を用いて, 有意水準 α をもつ非ランダム検定が作れるようになったので, ここからは検出力の比較に入る. まず, χ^2 適合度検定の検出力については本節の前半で述べたようにして求める. 一方, マックス検定の検出力は

$$\begin{aligned}\beta :=& P_K(S \ge c + 1 - \gamma) \\ =& P_K\left(\max_{1 \le i \le k} X_i \ge c + 1\right) \\ &+ \gamma\left\{P_K\left(\max_{1 \le i \le k} X_i \ge c\right) - P_K\left(\max_{1 \le i \le k} X_i \ge c + 1\right)\right\}\end{aligned}$$

となるので, 数値計算により検出力 β を求めることができる.

例 2 いま, $n = 1000$, $k = 12$, $\alpha = 0.05$ とし,

$$H : (p_1, \cdots, p_{12}) = (1/12, \cdots, 1/12)$$

とする. 例 1 で見たように

$$P_H\left(\max_{1 \le i \le 12} X_i > 106\right) = 0.0611$$
$$P_H\left(\max_{1 \le i \le 12} X_i > 107\right) = 0.0450$$

であるから, $c = 106$, $\gamma = 0.3111$ が求まる. ここで, 例えば対立仮説 K が

$$K : (p_1, \cdots, p_{12}) = (0.08, \cdots, 0.08, 0.1, 0.1)$$

とすると, χ^2 適合度検定の検出力は 0.4176, マックス検定の検出力は 0.4144 となりさほど変わらないといえる. しかし, 対立仮説が

$$K : (p_1, \cdots, p_{12}) = (0.08, \cdots, 0.08, 0.0933, 0.0933, 0.0933)$$

とすると, χ^2 適合度検定の検出力は 0.2460, マックス検定の検出力は 0.2060 となり, χ^2 適合度検定の検出力の方が高くなる. また, 対立仮説が

$$K : (p_1, \cdots, p_{12}) = (0.08, \cdots, 0.08, 0.083, 0.09, 0.107)$$

とすると, χ^2 検定の検出力は 0.4422, マックス検定の検出力は 0.5153 となり, マックス検定の検出力の方が高くなる. このように対立仮説によってどちらの検定の検出力が高くな

るかが変わってくるため, 幾つかの傾向をもつ対立仮説について2つの検出力を比較してみる. そこで, 対立仮説 $K:(p_1,\cdots,p_{12})=(p_1^{(n)},\cdots,p_{12}^{(n)})$ として

$K_1:(0.081,\cdots,0.081,0.109)$
$K_2:(0.079,\cdots,0.079,0.105,0.105)$
$K_3:(0.077,\cdots,0.077,0.1023,0.1023,0.1023)$
$K_4:(0.077,0.077,0.077,0.077,0.077,0.077,0.08967,\cdots,0.08967)$
$K_5:(0.08,\cdots,0.08,0.084,0.088,0.092,0.096)$
$K_6:(0.08,\cdots,0.08,0.082,0.086,0.092,0.100)$
$K_7:(0.08,\cdots,0.08,0.082,0.082,0.088,0.108)$

を考えると次の結果が得られる.

	χ^2 適合度検定の検出力	マックス検定の検出力
K_1	0.4511	0.5788
K_2	0.6855	0.6683
K_3	0.8146	0.6922
K_4	0.2974	0.1677
K_5	0.2135	0.1931
K_6	0.2778	0.2841
K_7	0.4537	0.5471

上記の結果から, $p_i^{(n)}$ $(i=1,\cdots,12)$ の中で一つでも比較的大きく他の値から離れている値がある場合はマックス検定の方が高い検出力を示し, $p_i^{(n)}$ $(i=1,\cdots,12)$ の間で値の離れ具合が小さい場合は χ^2 適合度検定の方が高い検出力を示していることがわかる. このことは, 両者の検定の性質を考えれば妥当に見える.

例 3 いま, $n=1000,\ k=20,\ \alpha=0.05$ とし,

$$H:(p_1,\cdots,p_{20})=(0.05,\cdots,0.05)$$

とする. このとき

$$P_H\left(\max_{1\le i\le 20} X_i>69\right)=0.0683,$$
$$P_H\left(\max_{1\le i\le 20} X_i>70\right)=0.0461$$

となるから, $c=69,\ \gamma=0.1750$ が求まる. ここで, 対立仮説 $K:(p_1,\cdots,p_{20})=(p_1^{(n)},\cdots,p_{20}^{(n)})$

として

$K_1 : (0.049, \cdots, 0.049, 0.069)$
$K_2 : (0.048, \cdots, 0.048, 0.068, 0.068)$
$K_3 : (0.048, \cdots, 0.048, 0.0613, 0.0613, 0.0613)$
$K_4 : (0.046, 0.046, 0.046, 0.046, 0.046, 0.046, 0.046, 0.046, 0.046, 0.046, 0.054, \cdots, 0.054)$
$K_5 : (0.048, \cdots, 0.048, 0.0483, 0.0489, 0.0499, 0.0512, 0.053, 0.0551, 0.0574, 0.0602)$
$K_6 : (0.048, \cdots, 0.048, 0.0481, 0.0481, 0.0484, 0.0489, 0.05, 0.0525, 0.058, 0.07)$

を考えると,次の結果が得られる.

	χ^2 適合度検定の検出力	マックス検定の検出力
K_1	0.3056	0.4466
K_2	0.5999	0.6326
K_3	0.3704	0.3364
K_4	0.2543	0.1347
K_5	0.1958	0.1766
K_6	0.4381	0.5214

この場合も, K_1, K_6 のように $p_i^{(n)}$ $(i = 1, \cdots, 20)$ の中で一つでも比較的大きく他の値から離れている値がある場合はマックス検定の方が高い検出力を示し, K_3, K_4, K_5 のように $p_i^{(n)}$ $(i = 1, \cdots, 20)$ の間で値の離れ具合が小さい場合は χ^2 適合度検定の方が高い検出力を示していることがわかる.

3　ランダムカイ2乗適合度検定

確率ベクトル $(X_1, \cdots, X_k)$ を多項分布 $M_k(n, p_1, \cdots, p_k)$ に従うとする. 第2節で述べたように χ^2 統計量は離散の値を取るので, $P_H(\chi^2 \geq c_\alpha) = \alpha$ となる c_α は必ずしも存在しない. また, χ^2 統計量を用いて検定を行う時には, その分布が χ^2 分布に収束することを用いた. ここでは, χ^2 分布に収束することを用いずに検定を行うために, 次のようなランダム化を行う. まず, $S_1, \cdots, S_k$ を独立に, いずれも指数分布 Exp(1) に従う確率変数とする. そして,

$$U_i := \frac{S_i}{S_1 + \cdots + S_k}, \quad i = 1, \cdots, k$$

とすると

$$\sum_{i=1}^{k} U_i = 1, \ U_k = 1 - \sum_{i=1}^{k-1} U_i$$

となる. ここで, $(U_1,\cdots,U_{k-1})$ の同時確率密度関数 (j.p.d.f.) を求める. $S_1,\cdots,S_k$ の j.p.d.f. は

$$f_{S_1,\cdots,S_k}(s_1,\cdots,s_k)=\begin{cases} e^{-(s_1+\cdots+s_k)} & (s_1,\cdots,s_k>0),\\ 0 \ (\text{その他}) \end{cases}$$

となる. そこで,

$$U_k=S_1+\cdots+S_k,$$
$$U_i=\frac{S_i}{S_1+\cdots+S_k},\quad i=1,\cdots,k-1$$

と変数変換すれば, ヤコビアンは u_k^{k-1} になるので

$$f_{U_1,\cdots,U_k}(u_1,\cdots,u_k)=e^{-u_k}u_k^{k-1}\quad (u_1,\cdots,u_k>0,\ \sum_{i=1}^{k-1}u_i\le 1)$$

となる. よって, $(U_1,\cdots,U_{k-1})$ の j.p.d.f. は

$$f_{U_1,\cdots,U_{k-1}}(u_1,\cdots,u_{k-1})$$
$$=\int_0^\infty e^{-u_k}u_k^{k-1}du_k=(k-1)!\quad (u_1,\cdots,u_{k-1}>0,\ \sum_{i=1}^{k-1}u_i\le 1)$$

となり, $(U_1,\cdots,U_{k-1})$ は Dirichlet(ディリクレ) 分布 $D(1,\cdots,1;1)$ に従っていることがわかる ([W62], [JK72]). また

$$f_{U_1}(u_1)=(k-1)!\int_0^{1-u_1}\int_0^{u_2}\cdots\int_0^{u_k-1}du_k\cdots du_2=k(1-u_1)^{k-2}\quad (0<u_1<1)$$

になり, U_1 はベータ分布 $Be(1,k-1)$ に従い, また $U_2,\cdots,U_{k-1}$ も同様にそれぞれベータ分布 $Be(1,k-1)$ に従っていることがわかる. よって, 各 U_i の平均は $1/k$ になる. このように $U_1,\cdots,U_k$ は連続の値を取るが独立ではないので, この意味では多項分布に対応しているものといえる. そこで,

$$\chi'^2:=\sum_{i=1}^{k}\frac{(X_i-np_i+U_i-(1/k))^2}{np_i}$$

として, χ^2 統計量のランダム化を行うことにする. そして, この統計量を用いて第 2 節で述べた検定について考えることにする. ここで, 統計量 χ'^2 をランダム χ^2 統計量と呼び, この統計量を用いた検定のことをランダム χ^2 適合度検定または単にランダム χ^2 検定と呼ぶことにする.

まず, χ'^2 のモーメントを求める.

$$Y_i:=X_i-np_i,$$
$$V_i:=U_i-\frac{1}{k},$$
$$Z_i:=\frac{1}{\sqrt{p_i}}(X_i-np_i+U_i-\frac{1}{k})=\frac{1}{\sqrt{p_i}}(Y_i+U_i)$$

とおくと

$$\chi'^2 = \frac{1}{n}\sum_{i=1}^{k} Z_i^2$$

となるから, χ'^2 のモーメントを求めるために, Y_i, V_i の次のモーメントを得る.

$$\begin{aligned}
E(Y_i) &= 0 \\
E(Y_i^2) &= np_iq_i \quad (q_i = 1 - p_i) \\
E(Y_i^3) &= np_iq_i(1-2p_i) \\
E(Y_i^4) &= np_iq_i(1-6p_iq_i) + 3n^2p_i^2q_i^2 \\
E(Y_i^6) &= np_iq_i - 30np_i^2q_i^2(q_i-p_i)^2 + 25n^2p_i^2q_i^2 - 130n^2p_i^3q_i^3 + 15n^3p_i^3q_i^3 \\
E(Y_iY_j) &= -np_ip_j \\
E(Y_i^2Y_j) &= -np_ip_j(q_i-p_i) \\
E(Y_i^2Y_j^2) &= -np_ip_j\{1-2(p_i+p_j)+6p_ip_j\} + n^2p_ip_j\{1-(p_i+p_j)+3p_ip_j\} \\
E(Y_i^3Y_j) &= -np_ip_j(1-6p_i+6p_i^2) - 3n^2p_i^2p_j(1-p_i) \\
E(Y_i^4Y_j^2) &= -np_ip_j(1-14p_i-2p_j+42p_ip_j+36p_i^2-144p_i^2p_j-24p_i^3+120p_i^3p_j) \\
&\quad + n^2p_ip_j(1-17p_i-p_j+41p_ip_j+42p_i^2-156p_i^2p_j-26p_i^3+130p_i^3p_j) \\
&\quad + 3n^3p_i^2p_j(1-2p_i-p_j+6p_ip_j+p_i^2-5p_i^2p_j) \\
E(Y_iY_jY_l) &= 2np_ip_jp_l \\
E(Y_i^2Y_jY_l) &= 2np_ip_jp_l(1-3p_i) - n^2p_ip_jp_l(1-3p_i) \\
E(Y_i^2Y_j^2Y_l^2) &= 2np_ip_jp_l\{1-3(p_i+p_j+p_l)+12(p_ip_j+p_ip_l+p_jp_l)-60p_ip_jp_l\} \\
&\quad + n^2p_ip_jp_l\{-3+7(p_i+p_j+p_l)-26(p_ip_j+p_ip_l+p_jp_l)+130p_ip_jp_l\} \\
&\quad + n^3p_ip_jp_l\{1-(p_i+p_j+p_l)+3(p_ip_j+p_ip_l+p_jp_l)-15p_ip_jp_l\} \\
E(V_i) &= 0 \\
E(V_i^2) &= \frac{k-1}{k^2(k+1)} \\
E(V_i^3) &= \frac{2(k-1)(k-2)}{k^3(k+1)(k+2)} \\
E(V_i^4) &= \frac{3(k-1)(3k^2-7k+6)}{k^4(k+1)(k+2)(k+3)} \\
E(V_i^6) &= \frac{5(k-1)(53k^4-238k^3+415k^2-326k+120)}{k^6(k+1)(k+2)(k+3)(k+4)(k+5)} \\
E(V_iV_j) &= -\frac{1}{k^2(k+1)} \\
E(V_i^2V_j) &= -\frac{2(k-2)}{k^3(k+1)(k+2)}
\end{aligned}$$

$$E(V_i^2V_j^2) = \frac{k^3 - 2k^2 + 15k - 18}{k^4(k+1)(k+2)(k+3)}$$
$$E(V_i^3V_j) = -\frac{3(3k^2 - 7k + 6)}{k^4(k+1)(k+2)(k+3)}$$
$$E(V_i^4V_j^2) = \frac{9k^5 - 55k^4 + 425k^3 - 1097k^2 + 1270k - 600}{k^6(k+1)(k+2)(k+3)(k+4)(k+5)}$$
$$E(V_iV_jV_l) = \frac{4}{k^3(k+1)(k+2)}$$
$$E(V_i^2V_jV_l) = -\frac{(k-3)(k-6)}{k^4(k+1)(k+2)(k+3)}$$
$$E(V_i^2V_j^2V_l^2) = \frac{k^5 - 3k^4 + 69k^3 - 273k^2 + 790k - 600}{k^6(k+1)(k+2)(k+3)(k+4)(k+5)}$$

そこで, これらを用いて χ'^2 のモーメントを求める. まず,

$$\begin{aligned}
E(\chi'^2) &= \frac{1}{n}\sum_{i=1}^{k} E(Z_i^2) \\
&= \frac{1}{n}\sum_{i=1}^{k} \frac{1}{p_i}\left\{E(Y_i^2) + 2E(Y_i)E(V_i) + E(V_i^2)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{k}\left\{n(1-p_i) + \frac{k-1}{k^2(k+1)p_i}\right\} \\
&= \sum_{i=1}^{k}(1-p_i) + \frac{k-1}{nk^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i} \\
&= k - 1 + \frac{k-1}{nk^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i} \qquad (11)
\end{aligned}$$

となる. 次に

$$\begin{aligned}
E(Z_i^4) &= \frac{1}{p_i^2}E(Y_i + V_i)^4 \\
&= \frac{nq_i(1 - 6p_iq_i)}{p_i} + 3n^2q_i^2 + \frac{6nq_i(k-1)}{p_ik^2(k+1)} + \frac{3(k-1)(3k^2-7k+6)}{p_i^2k^4(k+1)(k+2)(k+3)}, \\
E(Z_i^2Z_j^2) &= \frac{1}{p_ip_j}E[(Y_i+V_i)^2(Y_j+V_j)^2] \\
&= n\{-1 + 2(p_i+p_j) - 6p_ip_j\} + n^2\{1 - (p_i+p_j) + 3p_ip_j\} + \frac{4n}{k^2(k+1)} \\
&\quad + \frac{nq_i(k-1)}{p_jk^2(k+1)} + \frac{nq_j(k-1)}{p_ik^2(k+1)} + \frac{k^3 - 2k^2 + 15k - 18}{p_ip_jk^4(k+1)(k+2)(k+3)}
\end{aligned}$$

となり

$$E(\chi'^4) = \frac{1}{n^2}\left\{\sum_{i=1}^{k} E(Z_i^4) + \sum_{i\neq j}\sum E(Z_i^2 Z_j^2)\right\}$$

$$= k^2 - 1 + \frac{1}{n}\left\{-k^2 - 2k + 2 + \sum_{i=1}^{k}\frac{1}{p_i} + \frac{2(k-1)}{k^2}\sum_{i=1}^{k}\frac{1}{p_i}\right\}$$

$$+ \frac{1}{n^2}\left\{\frac{3(k-1)(3k^2-7k+6)}{k^4(k+1)(k+2)(k+3)}\sum_{i=1}^{k}\frac{1}{p_i^2} + \frac{k^3-2k^2+15k-18}{k^4(k+1)(k+2)(k+3)}\sum_{i\neq j}\sum\frac{1}{p_i p_j}\right\}$$

であり，さらに

$$\{E(\chi'^2)\}^2 = (k-1)^2 + \frac{2(k-1)^2}{nk^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i} + \frac{(k-1)^2}{n^2k^4(k+1)^2}\left(\sum_{i=1}^{k}\frac{1}{p_i}\right)^2$$

となるから

$$V(\chi'^2) = E(\chi'^4) - \{E(\chi'^2)\}^2$$

$$= 2k - 2 + \frac{1}{n}\left\{-k^2 - 2k + 2 + \sum_{i=1}^{k}\frac{1}{p_i} + \frac{4(k-1)}{k^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i}\right\}$$

$$+ \frac{1}{n^2}\left\{\frac{4(k-1)(2k^3-4k^2-k+6)}{k^4(k+1)^2(k+2)(k+3)}\sum_{i=1}^{k}\frac{1}{p_i^2}\right.$$

$$\left. - \frac{4(k^3-4k^2-k+6)}{k^4(k+1)^2(k+2)(k+3)}\sum_{i\neq j}\sum\frac{1}{p_i p_j}\right\} \qquad (12)$$

になる．さらに χ'^2 の 3 次キュムラントについては計算に必要な部分だけ述べておくと

$$\kappa_3(\chi'^2) := E[\{\chi'^2 - E(\chi'^2)\}^3]$$

$$= E(\chi'^6) - 3E(\chi'^2)E(\chi'^4) + 2\left\{E(\chi'^2)\right\}^3 \qquad (13)$$

であり，ここで

$$E(\chi'^6) = k^3 + 3k^2 - k - 3$$

$$+ \frac{1}{n}\left\{-3k^3 - 21k^2 - 102k + 104 + (19+3k)\sum_{i=1}^{k}\frac{1}{p_i} + \frac{9(k-1)(2k-1)}{k^2(k+1)}\right.$$

$$\left. + \frac{3(k-1)(k+4)}{k(k+1)}\sum_{i=1}^{k}\frac{1}{p_i} + \frac{9(k-1)}{k^2(k+1)}\left(\sum_{i\neq j}\sum\frac{p_i^2}{p_j} + \sum_{i\neq j\neq l}\sum\sum\frac{p_i p_j}{p_l}\right)\right\}$$

$$+\frac{1}{n^2}\left[2k^3+18k^2+28k-24-(22+3k)\sum_{i=1}^{k}\frac{1}{p_i}+\sum_{i=1}^{k}\frac{1}{p_i^2}-\frac{18(k-1)(2k-1)}{k^2(k+1)}\right.$$

$$+\frac{1}{k(k+1)}\sum_{i=1}^{k}\frac{1}{p_i}\left\{-3(k-1)(k+6)-\frac{3(k-1)(2k^3+5k^2+9k-18)}{k^3(k+2)(k+3)}\right\}$$

$$+\frac{1}{k(k+1)}\sum_{i=1}^{k}\frac{1}{p_i^2}\left\{\frac{15(k-1)}{k}+\frac{40(k-1)(k-2)}{k^2(k+2)}\right.$$

$$\left.+\frac{9(k-1)(k+4)(3k^2-7k+6)}{k^3(k+2)(k+3)}\right\}$$

$$+\frac{1}{k(k+1)}\sum_{i\neq j}\sum\frac{1}{p_ip_j}\left\{\frac{3(k-1)}{k}-\frac{24(k-2)}{k^2(k+2)}\right.$$

$$\left.+\frac{3(k+4)(k^3-2k^2+15k-18)}{k^3(k+2)(k+3)}\right\}$$

$$-\frac{18(k-1)}{k^2(k+1)}\left(\sum_{i\neq j}\sum\frac{p_i^2}{p_j}+\sum_{i\neq j\neq l}\sum\sum\frac{p_ip_j}{p_l}\right)$$

$$-\frac{9(k-1)(3k^2-7k+6)}{k^4(k+1)(k+2)(k+3)}\sum_{i\neq j}\sum\frac{p_i}{p_j^2}$$

$$\left.-\frac{3(k^3-2k^2+15k-18)}{k^4(k+1)(k+2)(k+3)}\sum_{i\neq j\neq l}\sum\sum\frac{p_i}{p_jp_l}\right]$$

$$+\frac{1}{n^3}\left\{\frac{5(k-1)(53k^4-238k^3+415k^2-326k+120)}{k^6(k+1)(k+2)(k+3)(k+4)(k+5)}\sum_{i=1}^{k}\frac{1}{p_i^3}\right.$$

$$+\frac{3(9k^5-55k^4+425k^3-1097k^2+1270k-600)}{k^6(k+1)(k+2)(k+3)(k+4)(k+5)}\sum_{i\neq j}\sum\frac{1}{p_i^2p_j}$$

$$\left.+\frac{k^5-3k^4+69k^3-273k^2+790k-600)}{k^6(k+1)(k+2)(k+3)(k+4)(k+5)}\sum_{i\neq j\neq l}\sum\sum\frac{1}{p_ip_jp_l}\right\},$$

$$E(\chi'^2)E(\chi'^4)=(k-1)(k^2-1)$$

$$+\frac{1}{n}\left[(k-1)\left\{-k^2-2k+2+\sum_{i=1}^{k}\frac{1}{p_i}+\frac{2(k-1)}{k^2}\sum_{i=1}^{k}\frac{1}{p_i}\right\}+\frac{(k-1)^2}{k^2}\sum_{i=1}^{k}\frac{1}{p_i}\right]$$

$$+\frac{1}{n^2}\left[\frac{3(k-1)^2(3k^2-7k+6)}{k^4(k+1)(k+2)(k+3)}\sum_{i=1}^{k}\frac{1}{p_i^2}\right.$$

$$+\frac{(k-1)(k^3-2k^2+15k-18)}{k^4(k+1)(k+2)(k+3)}\sum_{i\neq j}\sum\frac{1}{p_ip_j}$$

$$\left.+\frac{k-1}{k^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i}\left\{-k^2-2k+2+\sum_{i=1}^{k}\frac{1}{p_i}+\frac{2(k-1)}{k^2}\sum_{i=1}^{k}\frac{1}{p_i}\right\}\right]$$

$$+\frac{1}{n^3}\left[\frac{k-1}{k^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i}\left\{\frac{3(k-1)(3k^2-7k+6)}{k^4(k+1)(k+2)(k+3)}\sum_{i=1}^{k}\frac{1}{p_i^2}\right.\right.$$
$$\left.\left.+\frac{k^3-2k^2+15k-18}{k^4(k+1)(k+2)(k+3)}\sum_{i\neq j}\sum\frac{1}{p_ip_j}\right\}\right],$$

$$\{E(\chi'^2)\}^3=(k-1)^3+\frac{1}{n}\left\{\frac{3(k-1)^3}{k^2(k+1)}\sum_{i=1}^{k}\frac{1}{p_i}\right\}$$
$$+\frac{1}{n^2}\left\{\frac{3(k-1)^3}{k^4(k+1)^2}\left(\sum_{i=1}^{k}\frac{1}{p_i}\right)^2\right\}+\frac{1}{n^3}\left\{\frac{(k-1)^3}{k^6(k+1)^3}\left(\sum_{i=1}^{k}\frac{1}{p_i}\right)^3\right\}$$

となる.

次に, $P_H(\chi'^2>{\chi'_\alpha}^2)=\alpha$ となる棄却限界値 ${\chi'_\alpha}^2$ を漸近的に求める. いま,

$$f:=\frac{2\{E(\chi'^2)\}^2}{V(\chi'^2)},\ c:=\frac{E(\chi'^2)}{f},\ \delta:=\frac{\kappa_3(\chi'^2)}{8c^3f}$$

とし, $\chi^2_\alpha(f)$ を自由度 f の χ^2 分布の上側 $100\alpha\%$ 点とすると

$${\chi'_\alpha}^2\approx c\chi^2_\alpha(f)\left[1+\frac{\delta-1}{3(f+2)(f+4)}\left\{\chi^4_\alpha(f)-2(f+4)\chi^2_\alpha(f)+(f+2)(f+4)\right\}\right] \tag{14}$$

と近似される ([Ta75]). このことから, 有意水準 α を定めれば, この検定の棄却限界値を漸近的に求めることができる.

4　χ^2 適合度検定とランダム χ^2 適合度検定の比較

第3節で提案したランダム χ^2 検定がどの程度良いものかを調べるために, χ^2 統計量の漸近的な棄却限界値との比較を行ってみる. かつて, Katti[K73] は χ^2 統計量についての正確な棄却限界値の評価を表にしているので, その棄却限界値とランダム χ^2 検定の棄却限界値, 自由度 $k-1$ の χ^2 分布の上側 $100\alpha\%$ 点との相対差を数値的に評価する.

実際には, χ^2 統計量が離散的な値を取るため, [K73] では上側棄却限界値 u_α と下側棄却限界値 l_α とその時の上側確率が与えられているので, それらの値を線形補間し c_α を作る. そのようにして求めた c_α とランダム χ^2 統計量の分布の上側 $100\alpha\%$ 点, 自由度 $k-1$ の χ^2 分布の上側 $100\alpha\%$ 点との相対差を計算する. また, (14) において $\chi^2_\alpha(f)$ の値が必要になるが, 一般に f は整数とは限らないので, その場合の求め方については次の例の中で述べる.

例 4 いま, $n=25,\ k=10,\ \alpha=0.10$ とし,

$$H:(p_1,\cdots,p_{10})=(1/10,\cdots,1/10)$$

とする. このとき, [K73] から

$$u_\alpha = 15.40\ (\text{上側確率}\ 0.0858)$$
$$l_\alpha = 14.60\ (\text{上側確率}\ 0.1069)$$

となるので, 上側確率が 0.1000 となるように線形補間すれば

$$c_\alpha = 14.8616$$

となる. 次に, ランダム χ^2 検定の棄却限界値を求める. そこで, $n = 25,\ k = 10,\ p_i = 0.1(i = 1, \cdots, 10)$ を (11), (12), (13) に代入すると

$$E(\chi'^2) = 9.0327,\ V(\chi'^2) = 17.4112,\ \kappa_3(\chi'^2) = 47.3413$$

より,

$$f = 9.3721,\ c = 0.9638,\ \delta = 0.7053$$

となる. これらの数値と自由度 f の χ^2 分布の上側 10% 点 $\chi^2_{0.1}(f)$ を (14) に代入して χ'^2_α を求めたいのだが, $\chi^2_{0.1}(f)$ の自由度 f が整数値を取っていないため, χ^2 分布の表等から直接に値を求めることができない. そこで, この問題を解決するために, 次の 3 つのような方法 (i)〜(iii) が考えられる.

(i) 自由度 9 と 10 の χ^2 分布の上側 10% 点を線形補間したものを $\chi^2_{0.1}(f)$ として用いる. この時は

$$\chi^2_{0.1}(f) = 15.1688$$

となる.
(ii) Y を自由度 ν の χ^2 分布に従う確率変数とすると

$$(\sqrt{2Y} - \sqrt{2\nu - 1}) \xrightarrow{L} N(0,1) \quad (k \to \infty)$$

が成り立ち, その収束はかなり速いことが知られている ([A03] の pp.149, 157 参照). これより, ν を大きくとれば u_α を標準正規分布の上側 100α% 点として

$$P(\sqrt{2Y} - \sqrt{2\nu - 1} \geq u_\alpha) = P\left(Y \geq \frac{1}{2}(\sqrt{2\nu - 1} + u_\alpha)^2\right)$$

とできる. いま, $y_\alpha(\nu)$ を自由度 ν の χ^2 分布の上側 100α% 点を表すものとし, さらに

$$\widehat{y}_\alpha(\nu) := \frac{1}{2}(\sqrt{2\nu - 1} + u_\alpha)^2$$

とおく. ここで, 与えられた f に対して

$$\nu < f < \nu + 1,\ f = \nu + p\ (0 < p < 1)$$

となる ν と p をとる. そして,

$$\epsilon_1 := y_\alpha(\nu) - \widehat{y}_\alpha(\nu),$$
$$\epsilon_2 := y_\alpha(\nu+1) - \widehat{y}_\alpha(\nu+1),$$
$$y_\alpha^*(f) = y_\alpha^*(\nu+p) := \widehat{y}_\alpha(\nu+p) + (1-p)\epsilon_1 + p\epsilon_2$$

として, $y_\alpha^*(f)$ を自由度 f の χ^2 分布の上側 $100\alpha\%$ 点とする. この例では, $\nu = 9$, $p = 0.3721$ となり $\chi_{0.1}^2(f) = y_\alpha^*(f) = 15.1707$ となる.
(iii) 自由度 f の χ^2 分布が $\alpha = f/2$, $\beta = 2$ のガンマ分布 $\Gamma(f/2, 2)$ と同じになることを用いてガンマ分布の上側 $100\alpha\%$ 点 γ_α を求める. この例では, $\chi_{0.1}^2(f) = \gamma_\alpha = 15.1708$ となる.

上記の方法で, (i) は簡便であるが, あまり正確とは言えない. (ii) は1つの補間法で, かなり精度は良いが, ここでは, コンピュータソフト (エクセル) を用いて簡単に数値を得ることができるので, (iii) の考え方を採用することにする. いまの場合は, $\chi_{0.1}^2(f) = 15.1708$ となる. このとき, この数値を (14) に代入するとランダム χ^2 統計量の分布の上側 $100\alpha\%$ 点 ${\chi'_\alpha}^2 = 14.8434$ が求まるので, c_α との差は 0.0182 となり, その相対差は 0.00123 となる. 一方, 自由度 9 の χ^2 分布の上側 10% 点は 14.6837 なので, c_α との差は 0.1779 となり, その相対差は 0.01197 となる. よって, ランダム χ^2 統計量の分布の上側 $100\alpha\%$ 点の方が χ^2 分布のそれよりも望ましい値といえる.

各 $n = 5(5)25$, $k = 2(1)10$, $\alpha = 0.10$, 0.05, 0.01 について, 上の例のように相対差 $|c_\alpha - {\chi'_\alpha}^2|/c_\alpha$, $|c_\alpha - \chi_\alpha^2(k-1)|/c_\alpha$ の値を計算したものを下表として与える. これらの表から $\alpha = 0.10$ の場合はランダム χ^2 適合度検定の方が χ^2 適合度検定より良くなっていると考えられる. また, $\alpha = 0.05$, 0.01 の場合にはどちらの検定も相対差のバラツキがあってどちらが良いとも言えないように見える.

表 : χ^2 分布とランダム χ^2 統計量の分布のそれぞれの上側 $100\alpha\%$ 点 $\chi_\alpha^2(k-1)$ と ${\chi'_\alpha}^2$ の c_α に対する相対差

$\alpha = 0.10$

n	k	${\chi'_\alpha}^2$	c_α	$\chi_\alpha^2(k-1)$	$\lvert c_\alpha - {\chi'_\alpha}^2\rvert/c_\alpha$	$\lvert c_\alpha - \chi_\alpha^2(k-1)\rvert/c_\alpha$
5	3	6.2490	6.5914	4.6052	0.05195	0.30134
5	4	8.0679	7.5761	6.2514	0.06492	0.17485
5	5	9.5519	7.9750	7.7794	0.19773	0.02452
5	6	10.8757	9.9565	9.2363	0.09233	0.07233
5	7	12.1165	11.8078	10.6446	0.02614	0.09851
5	8	13.3115	13.4866	12.0170	0.01298	0.10897
5	9	14.4803	14.9602	13.3616	0.03208	0.10686
5	10	15.6334	16.4667	14.6837	0.05060	0.10828
10	2	3.1088	3.8994	2.7055	0.20276	0.30617
10	3	5.3447	4.9030	4.6052	0.09007	0.06075

表 (続)：χ^2 分布とランダム χ^2 統計量の分布のそれぞれの上側 $100\alpha\%$ 点 $\chi^2_\alpha(k-1)$ と ${\chi'_\alpha}^2$ の c_α に対する相対差

$\alpha = 0.10$

n	k	${\chi'_\alpha}^2$	c_α	$\chi^2_\alpha(k-1)$	$\lvert c_\alpha - {\chi'_\alpha}^2\rvert/c_\alpha$	$\lvert c_\alpha - \chi^2_\alpha(k-1)\rvert/c_\alpha$
10	4	7.0800	6.9276	6.2514	0.02200	0.09760
10	5	8.5925	7.9492	7.7794	0.08094	0.02135
10	6	9.9897	9.4560	9.2363	0.05644	0.02323
10	7	11.3202	11.0300	10.6446	0.02631	0.03494
10	8	12.6089	12.2748	12.0170	0.02722	0.02100
10	9	13.8697	13.6444	13.3616	0.01651	0.02073
10	10	15.1107	15.3279	14.6837	0.01417	0.04203
15	2	2.9601	3.7430	2.7055	0.20918	0.27718
15	3	5.0804	5.1129	4.6052	0.00636	0.09931
15	4	6.7866	6.5036	6.2514	0.04351	0.03878
15	5	8.3058	7.9247	7.7794	0.04809	0.01833
15	6	9.7244	9.3325	9.2363	0.04199	0.01031
15	7	11.0821	10.9014	10.6446	0.01658	0.02355
15	8	12.3998	12.3167	12.0170	0.00674	0.02433
15	9	13.6893	13.6263	13.3616	0.00462	0.01943
15	10	14.9580	14.9713	14.6837	0.00089	0.01921
20	2	2.8912	3.5727	2.7055	0.19076	0.24271
20	3	4.9550	4.9196	4.6052	0.00720	0.06391
20	4	6.6465	6.3039	6.2514	0.05434	0.00834
20	5	8.1684	7.9619	7.7794	0.02594	0.02292
20	6	9.5971	9.3722	9.2363	0.02400	0.01450
20	7	10.9679	10.8202	10.6446	0.01366	0.01622
20	8	12.2997	12.2301	12.0170	0.00569	0.01742
20	9	13.6033	13.5320	13.3616	0.00527	0.01260
20	10	14.8856	14.9011	14.6837	0.00104	0.01459
25	2	2.8515	3.4335	2.7055	0.16950	0.21201
25	3	4.8819	4.8256	4.6052	0.01167	0.04567
25	4	6.5644	6.3127	6.2514	0.03988	0.00971
25	5	8.0879	7.8535	7.7794	0.02985	0.00943
25	6	9.5224	9.3749	9.2363	0.01574	0.01477
25	7	10.9010	10.7989	10.6446	0.00946	0.01428
25	8	12.2411	12.1552	12.0170	0.00706	0.01137
25	9	13.5530	13.5039	13.3616	0.00363	0.01054
25	10	14.8434	14.8616	14.6837	0.00123	0.01197

表 (続) : χ^2 分布とランダム χ^2 統計量の分布のそれぞれの上側 $100\alpha\%$ 点 $\chi^2_\alpha(k-1)$ と ${\chi'_\alpha}^2$ の c_α に対する相対差

$\alpha = 0.05$

n	k	${\chi'_\alpha}^2$	c_α	$\chi^2_\alpha(k-1)$	$\lvert c_\alpha - {\chi'_\alpha}^2\rvert/c_\alpha$	$\lvert c_\alpha - \chi^2_\alpha(k-1)\rvert/c_\alpha$
5	3	8.1495	8.5347	5.9915	0.04514	0.29799
5	4	9.6781	9.9652	7.8147	0.02880	0.21580
5	5	10.9437	10.9750	9.4877	0.00286	0.13551
5	6	12.1265	11.6722	11.0705	0.03892	0.05155
5	7	13.2879	12.9287	12.5916	0.02778	0.02608
5	8	14.4498	15.0476	14.0671	0.03973	0.06516
5	9	15.6200	17.0679	15.5073	0.08483	0.09143
5	10	16.8006	18.9778	16.9190	0.11472	0.10849
10	2	4.7059	5.4922	3.8415	0.14317	0.30056
10	3	7.0136	6.4951	5.9915	0.07983	0.07754
10	4	8.7147	7.6833	7.8147	0.13424	0.01710
10	5	10.2046	9.3419	9.4877	0.09234	0.01561
10	6	11.6032	11.2776	11.0705	0.02888	0.01836
10	7	12.9565	12.9538	12.5916	0.00021	0.02796
10	8	14.2846	14.5878	14.0671	0.02078	0.03569
10	9	15.5971	15.9615	15.5073	0.02283	0.02846
10	10	16.8989	17.3375	16.9190	0.02530	0.02414
15	2	4.4003	5.0216	3.8415	0.12372	0.23501
15	3	6.6597	6.2303	5.9915	0.06892	0.03833
15	4	8.4068	7.7648	7.8147	0.08269	0.00643
15	5	9.9620	9.2514	9.4877	0.07681	0.02555
15	6	11.4252	11.2710	11.0705	0.01369	0.01779
15	7	12.8369	12.8402	12.5916	0.00026	0.01936
15	8	14.2160	14.2964	14.0671	0.00563	0.01604
15	9	15.5724	15.8600	15.5073	0.01813	0.02224
15	10	16.9120	17.3122	16.9190	0.02312	0.02271
20	2	4.2541	4.7905	3.8415	0.11199	0.19811
20	3	6.4876	6.2600	5.9915	0.03636	0.04290
20	4	8.2557	7.8182	7.8147	0.05596	0.00044
20	5	9.8419	9.5221	9.4877	0.03358	0.00361
20	6	11.3362	11.1087	11.0705	0.02048	0.00344
20	7	12.7761	12.7108	12.5916	0.00514	0.00938
20	8	14.1800	14.1775	14.0671	0.00018	0.00779
20	9	15.5579	15.7226	15.5073	0.01048	0.01370

表 (続) : χ^2 分布とランダム χ^2 統計量の分布のそれぞれの上側 $100\alpha\%$ 点 $\chi^2_\alpha(k-1)$ と ${\chi'_\alpha}^2$ の c_α に対する相対差

$\alpha = 0.05$

n	k	${\chi'_\alpha}^2$	c_α	$\chi^2_\alpha(k-1)$	$\lvert c_\alpha - {\chi'_\alpha}^2\rvert/c_\alpha$	$\lvert c_\alpha - \chi^2_\alpha(k-1)\rvert/c_\alpha$
20	10	16.9159	17.1280	16.9190	0.01239	0.01220
25	2	4.1684	4.6738	3.8415	0.10814	0.17809
25	3	6.3859	6.0585	5.9915	0.05405	0.01106
25	4	8.1660	7.8470	7.8147	0.04065	0.00411
25	5	9.7703	9.4877	9.4877	0.02979	0.00001
25	6	11.2829	11.2121	11.0705	0.00631	0.01264
25	7	12.7394	12.6906	12.5916	0.00385	0.00780
25	8	14.1580	14.2137	14.0671	0.00392	0.01031
25	9	15.5485	15.6743	15.5073	0.00802	0.01065
25	10	16.9174	17.1301	16.9190	0.01241	0.01232

$\alpha = 0.01$

n	k	${\chi'_\alpha}^2$	c_α	$\chi^2_\alpha(k-1)$	$\lvert c_\alpha - {\chi'_\alpha}^2\rvert/c_\alpha$	$\lvert c_\alpha - \chi^2_\alpha(k-1)\rvert/c_\alpha$
5	3	8.4406	8.1301	9.2104	0.03820	0.13287
5	4	8.6846	14.3338	11.3449	0.39412	0.20852
5	5	9.3400	17.9000	13.2767	0.47821	0.25828
5	6	10.3020	20.4238	15.0863	0.49559	0.26134
5	7	11.4745	21.3984	16.8119	0.46377	0.21434
5	8	12.7907	21.7509	18.4753	0.41194	0.15059
5	9	14.2063	23.3574	20.0902	0.39179	0.13988
5	10	15.6907	24.2000	21.6660	0.35162	0.10471
10	2	7.5994	8.5231	6.6349	0.10837	0.22154
10	3	9.0267	9.8261	9.2104	0.08135	0.06266
10	4	10.2904	11.8506	11.3449	0.13165	0.04267
10	5	11.6302	13.0769	13.2767	0.11063	0.01528
10	6	13.0462	15.7640	15.0863	0.17240	0.04299
10	7	14.5163	16.8211	16.8119	0.13702	0.00055
10	8	16.0214	18.7467	18.4753	0.14537	0.01447
10	9	17.5481	20.7080	20.0902	0.15259	0.02984
10	10	19.0874	22.9474	21.6660	0.16821	0.05584
15	2	7.2917	7.8203	6.6349	0.06759	0.15158
15	3	9.1283	9.3639	9.2104	0.02515	0.01639
15	4	10.6973	11.4000	11.3449	0.06164	0.00483
15	5	12.2433	13.1359	13.2767	0.06795	0.01072

表 (続) : χ^2 分布とランダム χ^2 統計量の分布のそれぞれの上側 $100\alpha\%$ 点 $\chi^2_\alpha(k-1)$ と ${\chi'_\alpha}^2$ の c_α に対する相対差

$\alpha = 0.01$

n	k	${\chi'_\alpha}^2$	c_α	$\chi^2_\alpha(k-1)$	$\lvert c_\alpha - {\chi'_\alpha}^2\rvert / c_\alpha$	$\lvert c_\alpha - \chi^2_\alpha(k-1)\rvert / c_\alpha$
15	6	13.7962	15.4857	15.0863	0.10910	0.02579
15	7	15.3552	16.9496	16.8119	0.09407	0.00813
15	8	16.9158	19.0091	18.4753	0.11012	0.02808
15	9	18.4738	20.7900	20.0902	0.11141	0.03366
15	10	20.0267	22.2083	21.6660	0.09823	0.02442
20	2	7.1323	7.7087	6.6349	0.07477	0.13930
20	3	9.1633	9.4375	9.2104	0.02905	0.02407
20	4	10.8792	11.0870	11.3449	0.01874	0.02326
20	5	12.5249	12.9778	13.2767	0.03490	0.02303
20	6	14.1439	15.0571	15.0863	0.06065	0.00194
20	7	15.7459	16.9906	16.8119	0.07326	0.01052
20	8	17.3330	18.7394	18.4753	0.07505	0.01409
20	9	18.9060	20.4786	20.0902	0.07679	0.01897
20	10	20.4651	22.2333	21.6660	0.07953	0.02552
25	2	7.0350	7.7413	6.6349	0.09124	0.14293
25	3	9.1796	8.9857	9.2104	0.02158	0.02500
25	4	10.9817	11.1067	11.3449	0.01125	0.02145
25	5	12.6861	13.2800	13.2767	0.04472	0.00025
25	6	14.3442	15.2277	15.0863	0.05802	0.00928
25	7	15.9715	16.9067	16.8119	0.05531	0.00561
25	8	17.5743	18.7376	18.4753	0.06208	0.01400
25	9	19.1560	20.4000	20.0902	0.06098	0.01519
25	10	20.7186	22.1200	21.6660	0.06335	0.02052

5 おわりに

本論の第 2 節では, まず, カイ 2 乗統計量の漸近分布等について簡単に述べ, そして, より簡便な検定統計量として, 最大値統計量を用いたマックス検定を考えて, カイ 2 乗適合度検定との比較を行った. 実際には, まず, マックス検定の第 1 種, 第 2 種の過誤の確率を [Lev81] の結果を用いて求め, マックス検定のランダム化を行うことによって有意水準をちょうど達成できるような検定とした. そして, マックス検定とカイ 2 乗適合度検定の検出力を比較し, 対立仮説の傾向によってどちらの検出力が高くなるかを調べた. その結果, 状況によってはマックス検定の方がカイ 2 乗適合度検定の方が検出力が大きくなることが分かった.

さらに，第3節では，カイ2乗統計量をランダム化することによって，連続の値をとることができるようにしたランダムカイ2乗統計量による検定を提案して論じた．実際，ランダムカイ2乗統計量による検定を考え，その分布の上側100α%点を近似的に求めた．そして，第4節において，その上側100α%点を[K73]によるカイ2乗適合度検定の棄却限界値を用いて，カイ2乗分布の上側100α%点と比較した．その結果，有意水準$\alpha = 0.10$の場合には，ランダムカイ2乗適合度検定の方がカイ2乗適合度検定よりも，与えられた水準に近い検定となることが分かった．

今後の課題としては，ランダムカイ2乗統計量を用いた検定の検出力を求め，カイ2乗適合度検定の検出力と比較してみることにより，ランダム化により検定を改良できたかどうかの判断材料が増えることになるであろう．

参考文献

[A03] 赤平　昌文 (2003). 統計解析入門. 森北出版.

[GN96] Greenwood, P. E. and Nikulin, M. S. (1996). *A Guide to Chi-squared Testing.* Wiley, New York.

[JK72] Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions.* Wiley, New York.

[JK92] Johnson, N. L. and Kotz, S. (1992). *Univariate discrete distributions.* Wiley, New York.

[K73] Katti, S. K. (1973). Exact distribution for the chi-square test in the one-way tables. *Commun. Statist.－Theory Meth.,* **2**, 435-447

[Leh99] Lehmann, E. L. (1999). *Elements of Large-Sample Theory.* Springer, New York.

[Lev81] Levin, B. (1981). A representation for the multinomial cumulative distribution function. *Ann. Statist.,* **9**, 1123-1126.

[Ta75] 竹内　啓 (1975). 確率分布の近似. 教育出版.

[To96] Torigoe, N. (1996). Approximations to non-central χ^2 and F distributions. *J. Japan Statist. Soc.* **26** (2), 145-159.

[W62] Wilks, S. S. (1962). *Mathematical Statistics.* (2nd ed.), Wiley, New York.